Quantifying isotopic heterogeneity of candidate reference materials at the picogram sampling scale

Article (Accepted Version)

Quantifying Isotopic Heterogeneity of Candidate Reference Materials at the Picogram Sampling Scale

Michael H. Ramsey[1] and Michael Wiedenbeck[2]

[1] School of Life Sciences, University of Sussex, Brighton, UK

m.h.ramsey@sussex.ac.uk

[2] Deutsches GeoForschungsZentrum GFZ, D-14473 Potsdam, Germany

Abstract

We propose a method for quantifying the *in situ* heterogeneity of solid materials at the picogram test portion scale, illustrating its use by investigating the oxygen isotope ratio ($^{18}O^-/^{16}O^-$) of four quartz samples. Using Secondary Ion Mass Spectrometry we could estimate the intrinsic heterogeneity using a large number (~100) of closely-spaced duplicated measurements. An analysis of variance was then applied to these large data sets to extract the measurement repeatability (typically 0.10 - 0.15‰, 1s) from the total variability, thereby revealing a variability ranging from 0.18 ‰ to 2.3 ‰ which can be attributed to the genuine isotope ratio heterogeneities. A small proportion of outlying values were either rejected manually after inspection, or were accommodated using robust statistics. We also evaluated two distinct approaches for estimating and correcting instrumental drift; the use of a sub-area of the test material (if shown to have sufficiently low heterogeneity) is judged to be preferable to using a piece of unrelated silicate glass which we believe to be homogeneous. We also compared three approaches for estimating measurement repeatability, from which we show that the 'duplicate method' applied to the reference material is preferable to using other methods based either on the drift monitoring material or on assessing residuals of the drift monitoring material after drift correction. Finally, here we propose a strategy for predicting the number of measurements on individual fragments of a material that would be required to achieve a specified target uncertainty.

Keywords

Heterogeneity, inhomogeneity, measurement repeatability, measurement uncertainty, oxygen isotopes, SIM S, microanalysis, calibration.

Introduction

Modern microanalytical technologies can provide isotope ratio determinations at sampling masses in the low nanogram (Laser Ablation Inductively Coupled Mass Spectrometry - LA-ICP-M S) to mid-picogram (Secondary Ion M ass Spectrometry - SIMS), with repeatabilities approaching or even surpassing 0.1 ‰ (1s) (e.g., Ashwal *et al.* 2017, $\delta^{18}O$ repeatability of 0.07 ‰ on n = 10 zircon measurements; Nasdala *et al.* 2016, $\delta^{18}O$ repeatability of 0.07 ‰ on n = 39 zircon measurements). Such fine-scale sampling is often crucial for understanding geochemical processes in complex natural materials; however, this fundamental advantage of *in situ* microanalysis is accompanied by a need for well characterized reference materials (RMs) that are fit- for-purpose when operating at micrometre sampling dimensions. Because ionization process within such sophisticated and expensive mass spectrometers are influenced by the nature of the matrix being analysed, it is required that calibration materials be closely matched to the composition and mineralogy of the materials being investigated (Eiler *et al.* 1997, Fabrega *et al.* 2017). Thus, the greatest performance constraint of such microanalytical techniques is commonly imposed by the availability of relatively inexpensive, matrix matched RM s. The widespread lack of well-characterised RM s is therefore the 'weakest link' limiting the realisation of the full potential of these technologies.

In this paper we investigate the SIM S determination of $\delta^{18}O$ in quartz to highlight calibration issues challenging geochemical microanalysis, though similar problems also beset other microanalytical technologies. Quartz was chosen because it is constant in both its major element composition (i.e. silicon and oxygen only) and in its crystallographic structure. For such work, it is common that the material being examined is embedded in an epoxy mount along with one or a few, randomly selected, mm-size to 100 µm-sized pieces of suitable RM s. In our examples using $\delta^{18}O$ determinations the RMs are not only needed for estimating the bias in the observed $^{18}O^-/^{16}O^-$ value (e.g. the instrumental mass fractionation, IMF) but they are also used to estimate measurement repeatability. In the case of modern SIMS instrumentation measurement repeatability, along with the uncertainty in the "bulk" determination of a RM's isotopic composition, are commonly the two largest contributors to the overall uncertainty budget. Thus a comprehensive understanding of

a RM 's heterogeneity is crucial both to understand the risk of significant isotopic variability of the RM at the mm-scale and also to avoid an overly pessimistic assessment of measurement repeatability due to significant variability at the 100's of µm- scale.

Many microanalytical RM characterization projects assume that the data variability observed during bulk characterization – in our case the $\delta^{18}O$ values determined by gas source mass spectrometry – is a reliable approximation of the material's overall heterogeneity (Eggins and Shelley 2002). Often fine-scale homogeneity testing has been limited to little more than conducting a series of microanalytical analyses and comparing the repeatability to some assumed benchmark for acceptable data quality. To our knowledge no previous microanalytical RM characterizations have attempted to quantify the true heterogeneity of a material, nor have they attempted to distinguish such heterogeneity from "analytical noise" inherent to the analytical design or the laboratory technology being employed. For example, in a recent characterization of $\delta^{18}O$ in two quartz RMs, the repeatability of the SIMS measurements was reported, from which the materials were described as homogeneous but without the value for the heterogeneity explicitly quantified (Seitz *et al.* 2016). Further characterizations of $\delta^{18}O$ have been in glasses (Hartley *et al.* 2012) and in carbonates as a paleoclimate tool considering the Mg/Fe matrix effects (Rollion-Bard and Marin-Carbonne 2011). Fitzsimons *et al.* (2000) presented a large data set of 3120 oxygen isotope ratio measurements in quartz and discussed their precision.

Quantification of Heterogeneity

Traditionally the objective for RM producers has been to achieve "homogeneity" of the analyte in the test material. Sample milling can be employed at bulk scales, such that the specified minimum test portion mass will inevitably contain a large number of particles. Through an averaging of thousands, or even tens-of-thousands of particles, it is possible to effectively "homogenise" a material that is actually quite variable at the single grain level. This is because the heterogeneity expressed as the sampling variance (i.e. square of standard deviation) is inversely

proportional to the mass of the test portion, when the particle size is small compared to the total size of the bulk sample (Gy 1979). It effectively becomes impossible to detect variability between test portions once this 'constitutional' heterogeneity becomes small relative to the measurement repeatability that a given analytical technique can deliver. A statistical test is usually applied to confirm that the between-bottle heterogeneity is not significantly greater than the within-bottle heterogeneity (Ellison 2015) for a specified minimum test portion mass.

For the case of *in situ* microanalytical techniques, such a milling strategy aimed at minimizing the effects of heterogeneity is not appropriate, as such instrumentation is designed specifically to quantify variations at the µm-scale. A more constructive approach is therefore to acknowledge that heterogeneity is ubiquitous and that it is something to be quantified and discussed explicitly. Ramsey *et al*. (2013) described a method for actually quantifying material heterogeneity at the centimetre to metre sampling scale provided by portable XRF devices. These authors duplicated *in situ* measurements of Pb concentration in soil at a variety of different scales from 0.2 m to 20 m at three Pb-contaminated land sites after which an Analysis of Variation (ANOVA) assessment was used to separate the measurement repeatability from the true *in situ* heterogeneity (i.e. the heterogeneity of the raw test material *in situ*, without grinding).

A somewhat similar approach has been applied at the microscopic scale to quantify the 'inhomogeneity' (i.e. heterogeneity) of over 70 analytes in eight NIST glasses, using probe techniques such as LA-ICP-MS and EPMA (Jochum *et al*. 2011). M easurement repeatability was determined by repeated analysis of presumed homogeneous glasses. Analytes with inhomogeneity estimates of less than 2% relative standard deviation (RSD) were defined as 'homogeneous'. The values for the inhomogeneity were used to calculat e a revised uncertainty on the certified analyte concentration values for three spot sizes (80, 45 and 25 µm diameters), with the corresponding estimated test portion mass (1.0, 0.1 and 0.02 µg).

Where a pure gas of known composition can be introduced into the analytical instrument, this can be used as way to estimate the measurement repeatability without any contribution from heterogeneity. This approach was used for the determination of the heterogeneity of $\delta^{18}O$ and $\delta^{13}C$ in calcite RM s in the mass range 0.2 to 10 µg by continuous-flow isotope ratio mass spectrometry (CF-IRMS) (Ishimura *et al*. 2008).

The homogeneity index (H), proposed by Boyd *et al.* (1967), has been used to compare the variance expected from counting statistics to the total variance observed. If the latter is significantly larger than the former variance ($H > 1$) then the heterogeneity can be considered significant. The application of this idea to EPMA data sets to assess the heterogeneity of micro-analytical RMs has more recently been described in general terms by Harries (2014), and for the specific case of olivine by Pankhurst *et al.* (2017). The use of the H value has been reported to provide advantages as applied to EPMA data sets when use of ANOVA is precluded because a replicate measurement at exactly the same location is ruled out due to the disruptive effect of the first measurement on the second. However, this approach relies on the assumption that the counting statistics give a reliable estimate of the measurement repeatability, as proposed by Fitzsimons *et al.* (2000). On the positive side, this approach enables an estimate to be made of the number of measurements required to detect a specific proportion of heterogeneity within an observed total variance.

This current investigation assesses whether a rigorous approach to heterogeneity quantification can be devised for microanalytical RMs where test portions are in the low nanogram to picogram test portion range. Our work, however, has some important differences from these earlier studies. In our case, which is based on SIM S $\delta^{18}O$ determinations, there are two reasons why the analyst should avoid estimating repeatability by solely conducting multiple analyses on a 'homogeneous' glass. Firstly, any such reference glass is unlikely to be matrix-mat ched to the mineral under investigation, in our case quartz, and so may well diverge significantly from the ionization behaviour of primary test material (Eiler *et al.* 1997). Secondly, without prior evidence it cannot be automatically assumed that the comparison glass is homogeneous for $\delta^{18}O$. A further difference to the previous "bulk" studies is that a reference material producer should wish to avoid designating a rather arbitrary threshold of acceptable heterogeneity, such as the value of 2% suggested by Jochum *et al.* (2011).

Three strategies exist for assessing heterogeneity. The first is to set an arbitrary threshold (e.g., 2% in trace element abundances, as already discussed), below which the analyte in that material could be described as having an acceptably small heterogeneity. The second approach would set the threshold value on an objectively determined fitness-for-purpose criterion; in the case of isotope ratio

determinations one might, for example, set the threshold for acceptable heterogeneity at a level well below the technically achievable repeatability of the target analytical method (i.e., substantially smaller than 0.10‰ (1s) in the case of $\delta^{18}O$ SIM S determinations). The third option would report the estimated value of the heterogeneity, which would then be propagated onto the total uncertainty budget for the analyte. The latter is not done routinely and it would still not necessarily include systematic effects.

In terms of fitness-for-purpose, a further issue is the mass of the RM that will be needed for an intended analytical application. For bulk methods, the RM producer should specify a minimum test portion mass that must be analysed in order that the certified value, and its uncertainty, will be valid (ISO 2006). For microbeam techniques, employing test portion masses in the nanogram or picogram range, the topic of test portion mass become more problematic. One approach could be for a characterization report to state separate uncertainty estimates for an analyte for both bulk methods and for data acquired at a specified microanalytical test portion mass. Furthermore, supposing that a proposed RM is to be distributed to different labs as mm-size fragments, homogeneity testing must consider both within-fragment and between-fragment variability -- roughly analogous to within-bottle and between-bottle heterogeneity testing of bulk characterization studies. Clearly, a microanalytical heterogeneity assessment (both within- and between-fragments) would need to test a sufficiently large number of grains such that subsequent users of the material can be advised how many fragments should be used so as to avoid significant bias from a bulk-determined value (i.e. in the case of $\delta^{18}O$, milligram determinations by gas source mass spectrometry). Fortunately, modern instrumentation is readily capable of providing such large data sets. In our example of $\delta^{18}O$ in quartz, SIMS is capable of assessing several hundred particles in a single night's run, with a total analytical time of around 3 minutes per determination.

The uncertainty of a RM 's certified value needs to be estimated accurately, primarily because it is a major component of the uncertainty of the subsequent measurements made on routine geochemical test materials (TMs). If the former is underestimated, then a hypothesis tested using routine samples can be accepted when, in truth, it should be rejected. This study focuses on accurately estimating the uncertainty of the

certified value, by including all relevant heterogeneity components, but this discussion also has implications for the wider estimation of measurement uncertainty.

Objectives

Here we have investigated the role played by both medium and fine-scale heterogeneity during the charact erisation of microanalytical RMs (µRMs), and how such heterogeneity should optimally be described. Specifically, our goals are:

1. To compare three different options for estimating measurement repeatability.
2. To devise a general procedure for quantifying (e.g. isotopic) heterogeneity at picogram scale using the 'duplicate method'. Specifically, we have applied ANOVA to estimate heterogeneity while taking into account the repeatability of our analytical measurements.
3. To compare the effectiveness of robust against classical ANOVA, and manual removal of outliers, in overcoming the disproportional effects of a small proportion of "measurement outliers" that occasionally impact such instrumental analyses (e.g., when sample conductivity is locally poor or when a minor scratch is present on the surface near the analysis point).
4. To develop a method for calculating the minimum number of measurements based on a minimum number of fragments which would be required to achieve a target uncertainty, for materials of known heterogeneity.

Methods

The measurement device used for this study was the Potsdam large geometry CAM ECA 1280-HR SIM S, which was used to determine $^{18}O^-/^{16}O^-$ ratios (i.e. $\delta^{18}O$) on four quartz materials. This analytical tool has previously achieved a measurement repeatability of 0.067 ‰ (1s, n = 10, Ashwal *et al.* 2017) or 0.092 ‰ (1s, n = 100, Nasdala *et al.* 2016) on silicates, which is reasonably typical of SIMS of this specification. Briefly, a 2.5 nA $^{133}Cs^+$ primary beam, operating in Gaussian mode, was focused to a ~5 µm diameter spot on the surface of the highly polished sample. A 10 x 10 µm raster was applied during the analysis sequence, thereby producing a flat-bottom crater and suppressing within-run drift in the total fractionation over the

course of the individual runs. A 90 s pre-sputtering was applied prior to each analysis so as to remove the high-purity gold coat, needed for electrical conductivity, and also to establish equilibrium sputtering conditions. Prior to the onset of data acquisition the secondary ion optics were aligned using automatic centring routines for the tool's field aperture (x and y directions), the contrast aperture (x direction only) and for the energy centring on the 50 eV wide energy window. Data were collected in static multi-collection mode, using a NRM field controller for maximum magnetic field stability, giving simultaneous ^{18}O and ^{16}O measurement for 20 x 4s at each pit. Ions were detected using the L2' Faraday cup for $^{16}O^-$ and the H2' for $^{18}O^-$, using 10^{10} and 10^{11} Ω resistors, respectively. A single analysis consisted of 20 integrations lasting 4 s each. The count rate on $^{16}O^-$ was typically 2.5 billion ions/s. Low energy, normal incidence electron flooding was used for all measurements in order to suppress sample charging over the course of the ~3 minute-long analytical run. The test portion mass for a single analysis was typically ~400 pg, based on a volume determined using white light profilometry, in conjunction with a density of quartz of $\rho = 2.65$ g/cm^3 (Fig. 1).

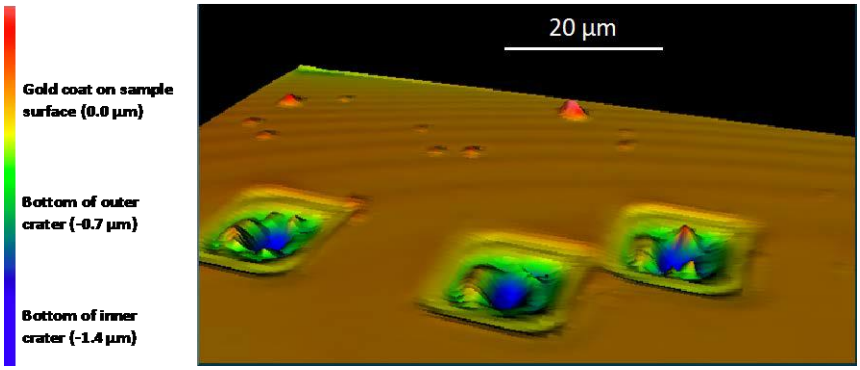


Figure 1

White light profilometer image of the surface of a grain of ZRM-1 quartz displaying thre e sputter crate rs, with a key to the colours used for the depth scale. The distance between the centres of the two left-most craters is 50 µm. The middle and right craters are an e xample of measurements that were made too close to each other geographically, disrupting the gold coating. This profilometer image determined a total volume for the ce ntral crater of ~120 µm^3, which in conjunction with a de nsity for quartz of 2.65 g/cm^3 yields a test portion mass of ~320 pg.

The Test Materials

The four quartz materials selected as candidate RMs for this study had no previous evidence of zoning in terms of oxygen isotopes, and were:

- NBS-28 – a high-purity silica sand, the starting material for which was provided by the Corning Glass Company, and subsequently washed in acid to remove impurities. The material had been sieved to a grain size fraction of 100 μm < ϕ < 177 μm prior to bottling in 0.5 g units (IAEA 2007). This material is available from the International Atomic Energy Agency, Vienna. It is the only material within this study that has been characterized for its oxygen isotopic composition, and has a recommended value of $\delta^{18}O_{VSMOW}$ = + 9.57 ± 0.10 ‰ (1 standard uncertainty), for which the reference sheet provides no guidance concerning minimum test portion mass (IAEA 2007). The data reported here all come from a single bottle of the material, from which we used several hundred individual grains to make our SIMS mount.
- ZRM -1 - high purity α-quartz with a certified SiO_2 content of >99.99 % (BAM 1991). This material was derived from a pegmatitic granite that was milled and then subjected to floatation, magnetic and electrostatic separation. The resulting quartz concentration was then cleaned using hydrofluoric and sulfuric acids. The mean grain size of the material, according to the certificat e of analysis, is ~150 μm. Our SIM S sample mount contained several hundred grains of ZRM-1 that were removed from a single, 100 g unit/bottle of the material as provided by BAM.
- GFZ-Qz1 - an 11 cm long synthetic quartz single crystal proved by the Korth Kritstalle GmbH, Altenholz, Germany that had an original mass of 133 g. This material is transparent and shows no visible inclusions; one side has a frosted appearance suggesting that material may have been sawed at an earlier time. We removed four corners by sawing from the ends of the original piece sent to Potsdam, and these were further sawed in order to produce four ~2-mm pieces that were used in this study.
- MfN-Qz1 - a 221 g single crystal of natural quartz bergkristall which is highly transparent and shows neither inclusions nor internal cracks. It shows a euhedral termination at one end. For our experiment four pieces were sawn off the corners of the crystal, and these pieces were further sawed to produce four

~2-mm pieces that were cast in epoxy. MfN-Qz1 was provided by the Museum für Naturkunde (MfN) in Berlin. Its history is not known, but it is likely to be of Alpine origin.

Thus we had two materials having fairly fine fragment sizes (NBS-28 and ZRM-1) and two that consisted of only four large fragments cut from larger single crystals (GFZ-Qz1 and MfN-Qz1). Each of these materials was made into its own individual sample block by casting in cold-set EpoFix two-component epoxy, which was ground and then polished to a flatness and roughness of better than 5 µm (Kita *et al.* 2009) as determined by white light profilometry (Fig 1). In the cases of NBS-28, ZRM-1 and MfN-Qz1, amm-size piece of NIST-610 glass was also cast in epoxy for each mount, which served as a monitor for instrumental stability. Interspersed analyses of Drift Monitors (DMs, of NIST-610 or quartz RM) were used to correct for a time dependent drift from sources such as instrumental mass fractionation, which for our study we believe results from steady variations in the gain of our Faraday Cup amplifier systems. Any such drift was presumed to be linear as a function of time, as suggested by Fitzsimons *et al.* (2000), who also suggest ed that the residual uncertainty of the regression be propagated into the SIMS analysis of unknowns.

Terminology and Nomenclature

The current VIM 3 definition of **measurement repeatability** is 'measurement precision under a set of repeatability conditions of measurement' (JCGM 2008). For this purpose, **repeatability condition of measurement** is defined as 'condition of measurement, out of a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and same location, and replicate measurements on the same or similar objects over a short period of time' (JCGM , 2008). This definition can be used to inform the choice of options for estimating measurement repeatability in this study. For example, 'replicate measurements on the same or similar objects' can be applied to measurements on adjacent areas of a crystal, where repeated measurements on the same exact spot are not feasible, due to technical reasons, such as the loss of gold coating, or the previous implantation of primary ions during the course of a prior measurement.

The duplicate measurements (say on a single grain of NBS-28) are therefore not really an 'analytical duplicate' but more a 'measurement duplicate' because they are based on slightly different samples. Measurement repeatability can therefore include a contribution from residual heterogeneity, even though in this study the heterogeneity at the 50 μm scale is initially assumed to be negligible. What is being estimated within these 'measurement duplicates' is therefore this 'measurement repeatability' (i.e. s_{meas}), and not the previously used term 'analytical repeatability', which is not discussed in the VIM 3 guide.

'Between-duplicates' variance is affected by any heterogeneity at the scale larger than the duplicate separation (e.g. 50 μm). If a RM consists of small fragments (e.g. NBS-28 with a grain size of ~230 um), the 'between-duplicates' variance gives an estimate of the overall heterogeneity at the scale between 50 μm and that of the bottle from which the RM is provided ($s_{hetero\ [50um - bottle]}$). For RMs provided in large fragments (e.g. GFZ-Qz1 ~ 2000 μm diameter), the variance 'between-duplicates' gives us the moderate-scale heterogeneity within the large fragments ($s_{hetero\ [50 - 2000\mu m]}$). However, we need to add the larger-scale heterogeneity between these large fragments ($s_{hetro\ [2000 - 100000\ \mu m]}$), to give the overall heterogeneity ($s_{hetero\ [50-100000\ \mu m]}$)

Experimental Design

Approximately 100 duplicated pairs (i.e. roughly 200 determinations) were measured for $^{18}O^-/^{16}O^-$ on the four candidate quartz RMs; the two members of these pairs were separated in space by 50 μm (or 60 μm for MfN-Qz1), but were randomized in their sequence within the analytical run (except for MfN-Qz1). For the fine-grained RMs, each duplicated pair was generally on a separate fragment, whereas for the course-grained RMs there were up to 30 duplicated pairs distributed quasi-randomly across each fragment.

Measurements on the drift monitoring area (whether glass or quartz) were also made as duplicate pairs with 50 μm spacing (60 μm for MfN-Qz1), but made sequentially. Such pairs of DM measurements were made at the beginning of each run, after every 5 pairs of measurements on the candidate RM, and finally at the end of each run. A single automated data acquisition sequence, including roughly 100 pairs of measurements along with the interspersed DM measurements, lasted around 15 hours.

Determining Repeatability and Heterogeneity

The simplest approach for estimating repeatability would be to make duplicated measurements on the same point of the test material, averaged over multiple points selected across the test material. However, SIMS is a destructive analytical method at the nanogram scale, so we could not repeat the second member of such a pair at the identical point to the first. Introducing a 50 (or 60) µm displacement between the members of the pair was found to be sufficient to maintain the sample's gold coat between the two locals and also to prevent any observable charging of the earlier analytical location during the subsequent analysis. By having this small distance between the duplicated measurements, our approach can employ the 'duplicate method' within an ANOVA-based data evaluation strategy. It also avoids the over-simplification of using the counting statistics alone to estimate the measurement repeatability, as proposed by Harries (2014) for what he referred to as analytical repeatability. Counting statistics, being a predicted variability assuming a theoretical data distribution, are not affected by any other sources of variability (such as drift), so will always give a minimum estimate of the repeatability which is not generally refelcted by an empirical approach. For SIMS such other potential sources in variability could include small variations in sample charging at the point of analysis, flickering of the flood electron source or small instabilities in the primary lens voltage supplies, etc. The risk that small-scale heterogeneity exists at the scale <50 µm, and so be included in estimates of repeatability, will be discussed below.

We have evaluated the repeatability of our measurements (s_{meas}) using three distinct methods:

(1) the within-duplicate variance of the DM measurements made on whatever DM material was being used for a given analytical run (before any drift correction was applied)

(2) the residual variance of the drift monitoring measurements after drift correction

(3) the within-duplicate variance of the main measurements on the given candidate RM (after drift correction, where drift is statistically significant).

The first method, based on duplicate consecutive measurements on the drift monitor in a closely space pair, will provide information on the short-term (few minutes) repeatability. This will be compared against the longer term repeatability (many hours) available from the second method using the drift monitoring measurements over the whole run (after drift correction), and the third method using the duplicated measurements on the RM within a randomomized sequence (except for MfN-Qz1).

The theoretical minimum number of duplicated measurements required, for perfectly mixed material having a Normal frequency distribution, has been estimated as eight (Lyn *et al*. 2007). Eight duplicate pairs will give what those authors consider an acceptable (but still non-negligible) confidence interval on the sampling variance arising mainly from a material's heterogeneity. However, for real world minerals with the possibility of unevenly distributed heterogeneity, and to further minimize this confidence interval, we recommend using a substantially larger number of duplicated measurements (and in some cases, the number of fragments). For the work reported here we aimed for around 100 pairs of analyses per candidate RM.

Application of Experimental Design to each of the Four Candidate RMs

The number and size of the sample fragments that were assessed varied between materials (Table 1). For NBS-28 and ZRM-1 the material was supplied as relatively small fragments, with mean lengths of the larger, often elongated crystals that we used for isotopic measurements of ~230 μm for NBS-28 and 280 μm for ZRM -1. M any such small fragments were selected at random from these materials. On each of the 100 fragments of NBS-28, one pair of duplicated measurements was made that were 50 μm apart; the 50 μm spacing was also used for the DM areas (Figure 2).

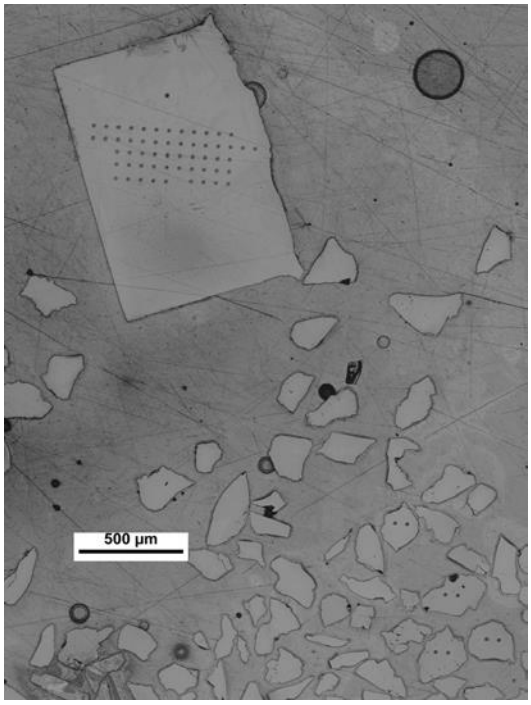


Figure 2. Reflected light photomicrograph of the NBS-28 mount, including a large r piece of NIST-610 glass, taken after the SIMS measurements. Duplicated measurement pits, used for the estimation of heterogeneity, can be seen 50 µm apart on several of the fragments of NSB-28 in the lower right of the image. A grid pattern of 55 pits, also with 50 µm spacing, used for drift monitoring, can be seen on the NIST-610 glass.

For ZRM -1 this design with 100 duplicates was applied to 70 fragments, of varying sizes (i.e., some of the fragments were large enough to accommodate 2 or even 3 duplicated measurements whilst aiming to retain the 50 µm spacing requirement, whereas 45 grains were measured with only a single duplicate pair). An additional sub-experiment for ZRM-1 aimed to investigate whether there was detectable heterogeneity within 8 of these larger fragments, when compared to the measurement repeatability determined at the 50 µm scale.

For the other two large quartz single crystal materials (GFZ-Qz1 & MfN-Qz1), four large fragments (~5000 µm) were removed from the corners of the given

crystal. Thirty duplicate sample pairs were measured on each of the four fragments of GFZ-Qz1, and 12 to 18 pairs for MfN-Qz1.

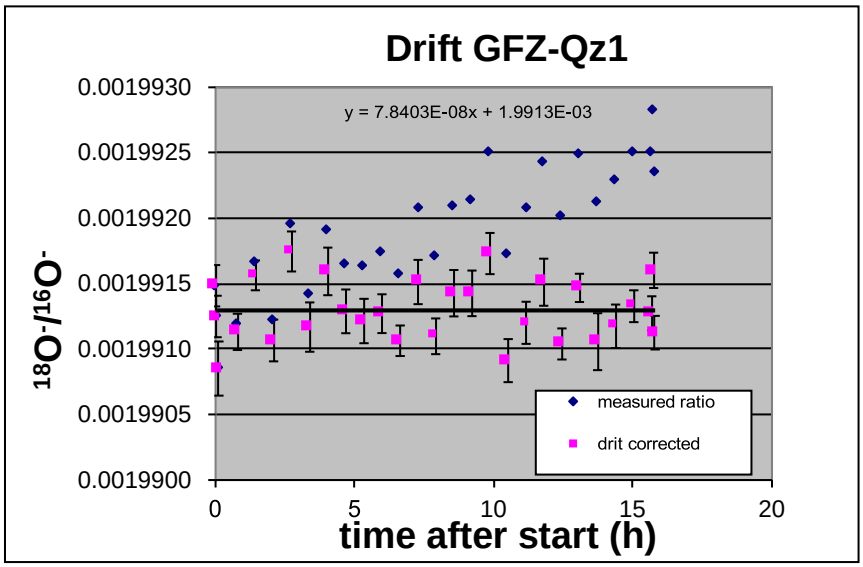
To achieve the best performance from the SIMS technique, it was essential to monitor the instrumental drift, by the general method already described above. Two approaches to achieving this were compared: (1) by making periodic measurements on a fragment of silicate glass (i.e. NIST-610) embedded in the mount with the test material (for NBS-28, ZRM -1 and MfN-Qz1), or (2) by making such DM measurements on a "small" sub-area of one of the larger fragments of the quartz test material itself (for GfZ-Qz1 & MfN-Qz1). For both approaches the area is initially assumed to be homogeneous for $\delta^{18}O$, but this assumption we tested subsequently. For MfN-Qz1 therefore, both options for drift monitoring were employed, thus allowing for a comparison between the two strategies.

Table 1. Instrumental parameters for measurements on four potential quartz RMs and their associated Drift Monitor, and number of outliers re jected.

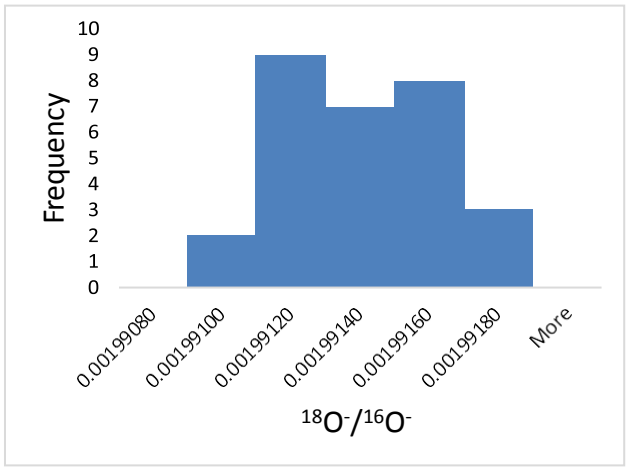
RM name	**Duration of run**	**Date of run start**	**Drift rate[a]**	**Number of RM fragments analysed**	**Number of Pairs** used in classical ANOVA (minus rejected pairs)		**Number of pairs rejected (with z-scores)[c]**
Material	**hours**	**day/month/year**	**$^{18}O^-/^{16}O^-$ /hour**	**n**	**RM**	**DM**	**RM**
NBS-28	14.5	02/07/2015	2.40 E-08	100	97	21[b]	3 (-4.8, -4.3, +3.3)
GFZ-Qz1	15.8	29/06/2015	7.84 E-08	4	120 (4*30)	14	0
ZRM-1	17.3	08/06/2015	3.62 E-08	70	90	26[b]	10 = 9[d] +1(+3.3)
MfN-Qz1	11.2	29/07/2016	0.77 E-08 (Not Sign.)	4	56 (10, 16, 12, 18)	12	2 (+4.1, +6.6)
			0.77 E-08[b] (Not Sign.)			14[b]	

[a]The drift rate is based upon the slope coeffi cient of the least-squares regression line of $^{18}O^-/^{16}O^-$ against time in hours. [b] Determined using NIST-610 silicate glass. [c] Rejection criterion usually based upon z-score of one measurement >3, [d] 9 pairs rejected due to overlap between duplicate craters.

Least-squares linear regression was used to model instrumental drift against time. Weighting was not applied as all points had similar variance. If the slope coefficient of the model was significantly different from zero (at 95% confidence), then the raw measurement values were corrected for this drift (Fig 3a). The residual variance after the correction, which was approximately normally distributed (as shown by the histogram Fig 3b), was also used in the second method for estimating measurement repeatability (s_{meas}).



(a)



(b)

Figure 3 (a). Example of the method of drift correction as applied to GFZ-Qz1, showing the e quation of the regression drift model fitted to the raw measurements (diamonds) and the same measurements after drift corre ctions (squares). Unce rtainty bars are instrumental repeatability ($1s_{inst}$), and only shown on corre cted measurements, for clarity. (b) The fre quency distribution of all 29 corrected measurements, showing an approximately normal distribution.

The *in situ* heterogeneity (s_{hetero}), and an estimate of measurement repeatability (s_{meas}) by the third method, were evaluated by applying ANOVA to the duplicated pairs of analyses on the RM being measured with random separation times, except for MfN-Qz1 (as opposed to the drift monitor duplicate measurements that were made sequentially).

Statistics-based identification of outlying measurement values.

The first criterion for the manual rejection of outlying measurement values requires the inspection of the location co-ordinates of each measurement point. If two craters were positioned appreciably less than 50 µm apart, visual inspection showed very erratic second measurement values which were ascribed to loss of the gold coating between the two craters (Fig 1). In this case, the second measurement value, and therefore the whole duplicate-pair to which that value belongs, was rejected. For example, 9 pairs of measurements were rejected for ZRM-Qz1 for this reason (Table 1, Fig. 1). The second criterion is the z-scores of any suspected measurement value (x), which compare it to the mean ($\bar{x}$) and standard deviation (s) of that set of measurements, using:

$$z = (x - \bar{x})/s$$

If the absolute z-score values are greater than 3, the measurement value should be rejected. For the example of NBS 28, the observed frequency distribution is shown in Fig 4

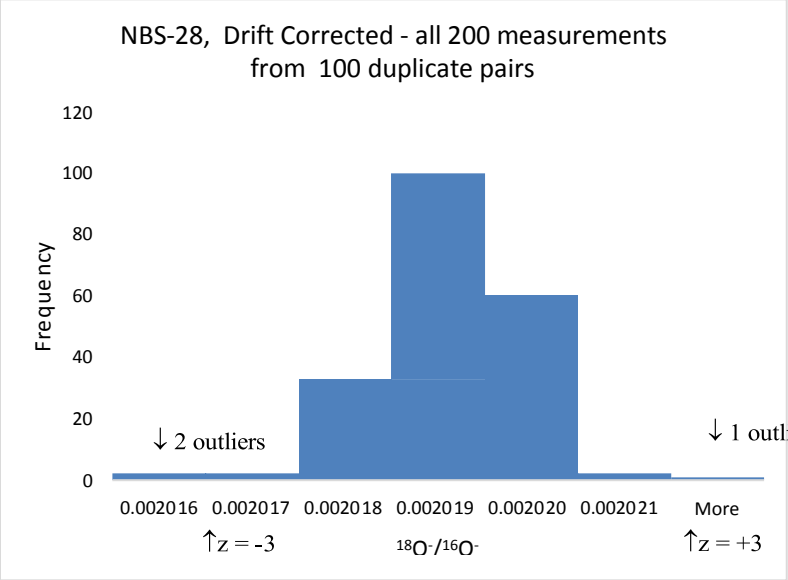


Figure 4. The fre quency distribution of all 200 measurements of $^{18}O^{-}/^{16}O^{-}$ on NBS-28 (drift corrected) with the upper-range limit shown. This chart shows a broadly normal distribution, but has three extreme values outside an z-score limit of ± 3, which are therefore considered outlying values.

Three of the 200 measurement values have absolute z-scores great er than 3, (at -4.8, -4.3 and +3.3), and are therefore more than 3s from the mean value. Assuming the frequency distribution of the measurement uncertainty is normal (Thompson and Howarth 1980), that model predicts that only 0.27% of the measurement values (i.e. 0.5 out of 200) should exceed this threshold, and only 0.006% should exceed a z-score of 4 (i.e. 0.01 out of 200). The fact that we have 3 such values for NBS-28, suggests that they are not members of the parent population, and can be reliably rejected. Each outlier was part of a different duplicate pair, so the number of pairs used in the ANOVA for NBS-28 was reduced from 100 to 97 pairs (Table 1). Other methods of detecting outlying values, such as Dixon's Q-test, were not applicable in this case, due to the very high number of observations available in the population for testing these values. A comparison between this approach and the use of robust statistics will be presented below.

Here we note that after the identification of outlier values using this statistical approach it is valuable to investigate the possible origins of such divergent results.

We found it useful to check the location on the sample mount which produced significantly divergent results. In many cases it was possible to attribute the outlier result to either surface problems (e.g., crater positioned on top of a scratch or crater within a few 100 µm of a major surface flaw) or to a sample specific problem (e.g., raster extending out onto the epoxy embedding media).

Application of ANOVA

This statistical analysis, and the comparison between the use of robust and classical ANOVA, was made using the software RANOVA (Version 1.0), (Analytical Methods Committee 2014; Rostron and Ramsey 2012). The detailed procedure of how this software was adapted for this purposeis described in Electronic Supplement A1. In descriptive terms, the ANOVA calculates both the total variance of all of the measurements being considered (s^2_{total}), and its component parts as shown in Equation 1. These components arise from the repeatability of each measurement value (s_{meas}), quantified (for Methods 1 & 3) from the variability *within* the duplicate pairs, and the heterogeneity (s_{hetero}), quantified from the variability *between* the different duplicate pairs.

$$s^2_{total} = s^2_{meas} + s^2_{hetero} \qquad (1)$$

For the fine-grained RMs, the heterogeneity is that for all scales from the 50 µm (used in the duplicate measurements) up to that of the bottle in which the RM was supplied (i.e. $s_{hetero[50\mu m\text{-}bottle]}$).

For the large-grained RMs, there are two components to the heterogeneity variance:

$$s^2_{hetero} = s^2_{hetero[within\text{-}frag]} + s^2_{hetero[between\text{-}frag]}$$

The application of ANOVA in this experimental design only gives the heterogeneity within each large fragment (i.e. $s_{hetero[50\mu m\text{-}within\text{-}frag]}$). The extra heterogeneity between the large fragments was estimated by subtracting total within-fragment variance from the total variance across all fragments, using Equation 2.

$$s_{hetro|between-frag|} = \sqrt{s^2_{total\text{-}all\ frags} - s^2_{within-frag}} \qquad (2)$$

For these large-grained RMs, the two component heterogeneity variances (within- and between-fragments) can then be added to give an estimate of the heterogeneity at all the scales from 50 µm up to the whole crystal (i.e. $s_{hetero[50\mu m\text{-}crystal]}$) using Equation 3.

$$s_{hetero[50\mu m\text{-}crystal]} = \sqrt{s^2_{hetero[50\mu m\text{-}within\text{-}frag]} + s^2_{hetro[between\text{-}frag]}} \qquad (3)$$

Robust statistical techniques have been devised as alternative to classical statistics (Huber 1989) and these have proved to be very effective for the interpretation of analytical measurements that have a quasi-Gaussian distribution, but which also contain a small proportion of outlying values (AMC 1989).

Our first and third methods for estimating measurement repeatability therefore used both classical and robust (R) ANOVA. The RANOVA program in particular was designed to accommodate up to 10% of measurement values beyond those belonging to the normal distribution assumed by classical ANOVA. The RANOVA software identifies values outside 1.5 standard deviations from the mean value of an initial classical estimation, and these values are 'accommodat ed' to equal 1.5s in a process called windsorisation. Iterative recalculation and adjustment of these two statistics then generat es results in the values of the robust mean and the robust standard deviation of the underlying normal distribution (Analytical M ethods Committee 2001). When applied to populations where there are no outlying values, the robust and classical statistics are the same.

Our third method for estimating measurement repeatability is based on applying ANOVA to the duplicated measurements of the candidate RM itself (rather than on the DM area used in the first method). This third method assumes that within the test material there is no appreciable heterogeneity of the analyte at the scale at the 50 or 60 µm scale which we used. The validity of this assumption was checked by comparing this value of s_{meas} with those made using the other two methods. For the large grain size materials (i.e. GFZ-Qz1 & MfN-Qz1) thepotential extra

heterogeneity between-fragments was added to establish the overall heterogeneity ($s_{hetero[50\mu m\text{-}crystal]}$) using Equation 3. As a test for possible effects related to matrix sputtering behaviour, in the case of MfN-Qz1, weran both NIST-610 silicate glass and a small area (~300 µm in size) of MfN-Qz1 as quasi-independent drift monitors.

The unit used to express both heterogeneity and measurement repeatability in this study is 'per mil' ‰ (i.e. $1000 * \frac{s}{\bar{x}}$), where *s* is the standard deviation and $\bar{x}$ is the mean value of the individual measurement values. This nomenclature was chosen as it is the unit traditionally used to express repeatability in oxygen isotope geochemistry, and also because it enables a direct comparison to be made between repeatability and heterogeneity. Previous studies have used RSD% (i.e. $100 * \frac{s}{\bar{x}}$) or the heterogeneity factor (i.e. HF = exp(s_G) where s_G is the standard deviation of the $\log_e$-transformed distribution of the measurement values) (Ramsey *et al*. 2013).

One component of measurement repeatability that can also be usefully compared is the instrumental repeatability (s_{inst}), which is called "internal uncertainty" in the software of this particular SIMS instrument. This is calculat ed as 1 standard error on the mean (s/√n) of the 20 time-contiguous isotope ratio measurements (each lasting 4 seconds), stretching over the total measurement period of 80 sec.

Results

NBS-28

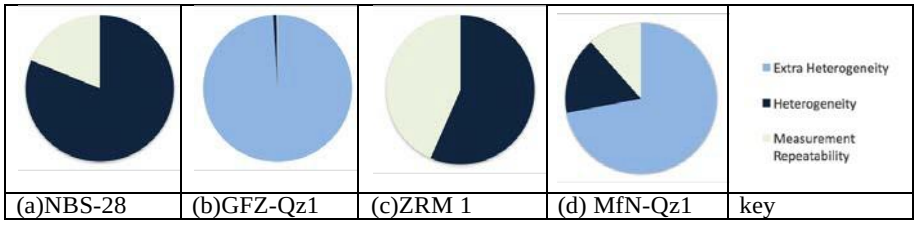
The 'total' repeatability on the fine-grained materials of NBS-28 is 0.31‰ (s_{total} in Table 2, from 194 measurements). The components that make up this variance, quantified using ANOVA, provide important information about the main sources of the observed variability.

Table 2. Summary of sources of variance from classical ANOVA for all four candidate RMs and their associated drift monitoring materials (DMs), expressed as ‰ for the number of measurements (drift corre cted in those cases where drift was significant), given in Table 1.

		Heterogeneity		Repeatability			
				Measurement[a]			Instru-mental (mean value)
RM	'Total' overall variance (‰)	Across fragments (including within-fragment)	Between-duplicates ([c] within-fragment)	Within DM duplicates	DM overall, post correction	Within-RM duplicates	
Symbol	*Stotal*	*Shetero[50μm-bottle] or [50μm-crystal]*	*Shetero[50μm-bottle] or [c][50μm-frag]]*	*Method1 Smeas*	*Method 2 Smeas*	*Method 3 Smeas*	*Sinst*
NBS-28	0.31	0.28	0.28	0.12[b]	0.12[b]	0.14	0.081
GFZ-Qz1	2.30	2.30	0.20[c]	0.13	0.12	0.10 [c]	0.078
ZRM-1	0.23	0.18	0.18	0.11[b]	0.11[b]	0.15	0.084
MfN-Qz1	0.27	0.25	0.12[c]	0.09 0.08[b]	0.10 0.10 [b]	0.10[c]	0.081

[a] Include s the instrumental repeatability. [b] Determined using NIST-610 glass, [c] Mean value of those calculated for each of the 4 fragments.

The relative contributions of all of these components of the total measurement variance on any test material can usefully be appreciated using a pie diagram (e.g. Fig 5).



(a)NBS-28	(b)GFZ-Qz1	(c)ZRM 1	(d) MfN-Qz1	key

Figure 5. Pie diagrams showing the proportions of the total variance of measurements for each of the four test materials, as the measurement repeatability (within-duplicate RM measurements by Method 3, s^2_{meas}, light tone), heterogeneity (between-duplicate measurements, $s^2_{hetero[within-frag]}$, dark tone), and for the cases (b) and (d) also the extra heterogeneity (between larger fragments of the test material, $s^2_{hetero[between-frag]}$, me dium tone)

The measurement repeatability for NBS-28, when estimated using Method 3 (with 97 duplicated measurements) is 0.14 ‰, which accounts for only 19% of the total

variance (Fig 5a). A lower estimate of this same component of 0.12 ‰ was obtained using a small area of NIST-610 glass DM, by both Methods 1 and 2. This difference between the two estimates of measurement repeatability may be due to an inherently poorer repeatability when analysing a quartz matrix as compared to a glass, or it may suggest a lower heterogeneity of $\delta^{18}O$ in the glass. It would not seem to be due to the time period over which the repeatability is estimated, as the longest duration (i.e. the whole run of over ten hours) in Method 2 gives the lower value of 0.12 ‰. One sub-component of the measurement repeatability is that arising from the instrumental (within-run) repeatability (mean value 0.08 ‰). This mean s_{inst} value does not vary significantly between the four quartz materials (NBS-28, ZRM 1, GFZ-Qz1 and MfN-Qz1 are 0.081, 0.078, 0.084 and 0.081 ‰ respectively), nor is it different for the glass (NIST-610 has 0.082 ‰). However, this accounts for only 33% of measurement repeatability ($100*(0.08^2/0.14^2)$) and 7% of the total variance. Other possible components that may account for the remaining 77%, are discussed below in the section 'Components of measurement repeatability'.

The heterogeneity ($s_{hetero[50\mu m\text{-}bottle]}$), estimated from the between-duplicate variance on the NBS-28 quartz, is 0.28 ‰ (Table 2). This value is less than the 'total' repeatability of 0.31‰, but nonetheless accounts for the main proportion (i.e. 81%) of that variance (Fig 5a).

GFZ-Qz1

For this synthetic single crystal the ANOVA results on the RM duplicates (Method 3) show a better than typical measurement repeatability of 0.10 ‰, but this does vary between the four fragments from 0.076 ‰ for F1, 0.094 ‰ for F3, 0.111 ‰ for F4, to the more typical 0.128 ‰ for F2. The estimates of the measurement repeatability from ANOVA of the DM duplicates on the quartz of this RM (Method 1) and the DM corrected results (Method 2) are similar at 0.13 ‰ and 0.12 ‰, respectively (Table 2). This was made using a sub-area (250 x 250 μm) of one fragment (F4) of GFZ-Qz1, rather than using the NIST-610 glass, as was the case for the other materials. It was serendipitous that F4 was chosen for this purpose, as this has the lowest heterogeneity of the four fragments (0.05 ‰, see below), compared with 0.23 ‰, 0.35 ‰ and 0.18 ‰ for F1, F2 and F3 respectively. The heterogeneity of the DM area within F4 is also very low, with drift-corrected $^{18}O^-/^{16}O^-$ measurements all being in the range 0.0019915 to 0.0019919. It seems be that using Method 3 gives a more meaningful estimate of

the measurement repeatability, as it uses duplicates placed on all four fragments, rather than just on the one fragment used for the DM.

The average total variance (including heterogeneity) within each fragment of GFZ-Qz1 is 0.24 ‰ [i.e. $\sqrt{(0.20^2 + 0.10^{2)}}$], which is smaller than was the case for NBS-28 (Table 2). However, this material was provided as a single, large crystal, of which four larger fragments (~5 mm) derived from distant parts of the crystal were used for our experiment. This material therefore has the potential to show an extra source of variance from the between-fragment heterogeneity, and this form of heterogeneity was estimated as the very large value of 2.30 ‰. The total repeatability is therefore 2.3 ‰ (i.e. equal to the between-fragment value after rounding). The heterogeneity contribution within each of the four fragments is quite low at 0.20 ‰ on average, but there is great variability in the heterogeneity across the fragments, with fragment 4 (F4 = 0.05 ‰) evidently being much less heterogeneous than fragment 2 (F2 = 0.35 ‰). This pattern is readily observed by plotting analytical positions against the determined $^{18}O^-/^{16}O^-$ values (Fig 6).

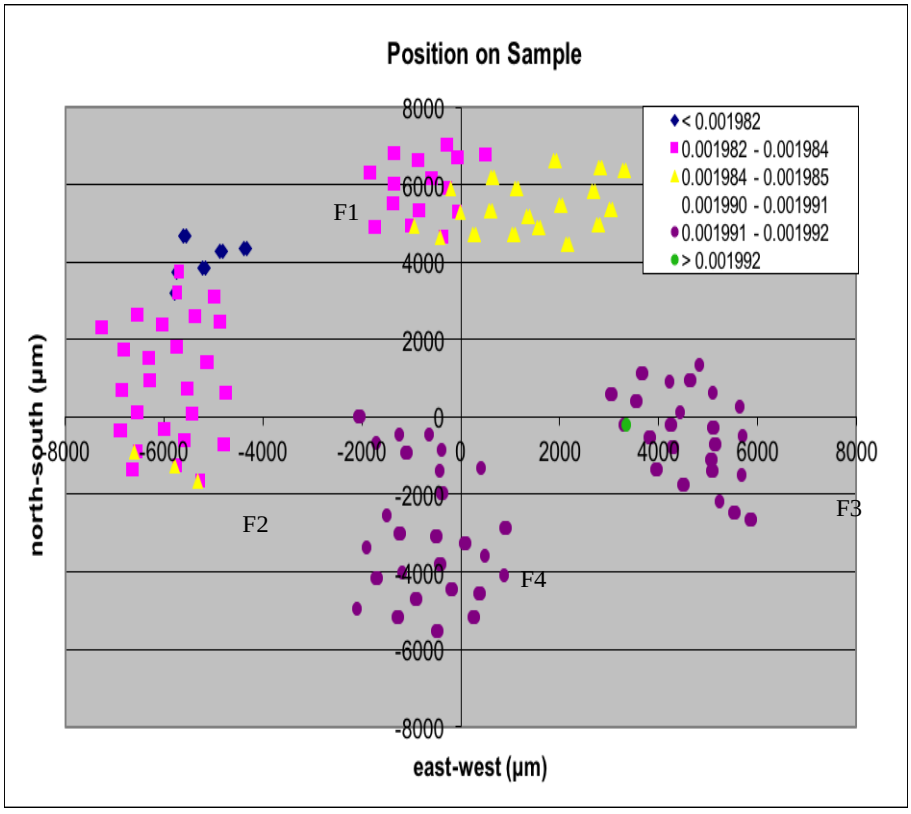


Figure 6. Distribution map of $^{18}O^{-}/^{16}O^{-}$ values for the four fragments of GFZ-Qz1. It shows high leve ls of heterogeneity both between the fragments, and within some of them (e.g. F2 on the western side), which makes this material unsuitable for de veloping as a $\delta^{18}O$ RM.

ZRM-1

The results from the fine-grained material ZRM -1 (Table 2) are based upon the application of classical ANOVA after the manual exclusion of 10 duplicate pairs (nine spots due to visibly overlapping craters, the result of randomly located pairs on single grains being too close, and one from high z-score of 3.3, Table 1). The raw measurements are particularly useful for illustrating the impact of outlying measurement results, and how they can best be dealt with (see Discussion).

The heterogeneity value estimated for ZRM 1 ($s_{hetero[50um-bottle]}$ = 0.18 ‰) is the lowest found for any of the four materials tested, which makes this a potentially

useful material for certification as a certified RM for $\delta^{18}O$. It has the lowest proportion of total variance contributed by the heterogeneity (60% - see Fig. 5c), but this may be partially caused by the larger measurement repeatability estimate of 0.15 ‰ (determined by Method 3, Table 2). The later may have been caused by either some heterogeneity at the 50 um scale, or some uncorrected drift between the duplicate RM measurements which had random time separations. The estimates made using Methods 1 & 2 are smaller (i.e. 0.11 ‰), which may be due to the use of the glass NBS-610 for the DM in Methods 1 & 2, but which bring with them the substantial disadvantage of not being matrix matched to the quartz test material. The sub-experiment to test whether there was detectable extra heterogeneity within eight larger fragments of ZRM-1 (i.e. between multiple duplicate pairs within these fragments) did not detect any.

MfN-Qz1

The average measurement repeatability of 0.10 ‰ for the single crystal MfN-Qz1, as estimated from the within-duplicate variance of the RM (Method 3, Table 2), is one of the smallest we observed. This estimate may be lower than those for two of the other RMs because the analytical duplicates for this RM were measured sequentially, rather than at random times within the run. However, this value is similar to those made using the DM approach (Method 2), using both the quartz in the RM, and the NIST-610 glass, even though no significant drift was detected for either material. The values estimated by Method 1 were also very similar when using either quartz (0.09 ‰) or glass (0.08 ‰) DMs, and were both lower than for any other values in these experiments. This is probably due to the fortuitously good machine stability during this run, including the absence of any detectable drift (Table 1).

The heterogeneity estimated within these large fragments (~5 mm) was the smallest for all RMs (average 0.12 ‰) and varied relatively little across the four fragments investigated (0.09 to 0.17 ‰). However, extra heterogeneity was detected between these four large fragments (0.25 ‰), but the overall heterogeneity is still the second lowest of the four materials tested. Shown as a proportion of the total variance (Fig. 5d), this extra heterogeneity seems almost as large as that for GFZ-Qz1 (86% to ~99%), but in absolute terms it is much smaller (0.25 ‰ vs. 2.30 ‰, see Table 2), and may be considered acceptable.

Discussion

Best method for estimating measurement repeatability

For the SIMS analytical method one uses reference materials for both instrument calibration as well as for estimating the overall repeatability of the analytical design. Such repeatability estimates (and calibrations) are based on the assumption that the RM material is effectively homogeneous in its composition. As we have seen above, in all four cases the quartz materials that we investigated this seems not to be the case. Here we develop a means of estimating measurement repeatability (s_{meas}) even when a modest level of heterogeneity might be present in a reference material.

Initial inspection of the values of measurement repeatability in Table 2 might suggest that Method 1 (duplicates on the DM material) gave the most consistent estimates (0.08 – 0.13 ‰). We note here that in all case M ethods 1 & 2 resulted in nearly identical results, with Method 2 providing the advantage of compensating for any drift in the experiment al conditions should such be present. As already mentioned, three of these values were made on the presumably homogeneous glass NBS 610, which is not matrix-mat ched to the quartz test materials. When a small area of the quartz RM itself was used for comparison in the MfN-Qz1 mount, the value was practically indistinguishable from that from the glass (0.09 ‰ or 0.10 ‰ for Methods 1 and 2, respectively; Table 2). Method 3 overcomes this problem of matrix matching by providing duplicate measurements on the actual matrix being tested. This strength must be set against the limitation that Method 3 is susceptible to heterogeneity of the analyte in the test material at the 50 μm scale. One solution maybe to use Method 3, but only when the test material has been demonstrated as having a sufficiently low level of overall heterogeneity, as should be the case of a high quality RM in any case.

The components of the measurement repeatability

One underlying and limiting factor in the measurement repeatability is the instrumental 'uncertainty' (s_{inst}). The value of this was remarkably constant between the runs for the four RMs, at around 0.08 ‰ (Table 2).

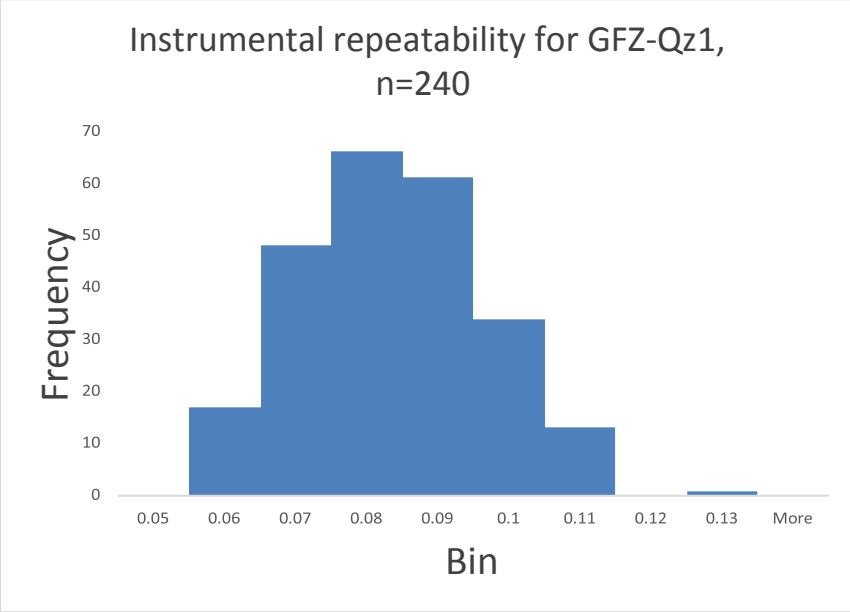


Figure 7. Fre quency distribution of estimates of the instrumental 'uncertainty' (s_{inst}) for 240 isotope ratio measurements on GFZ-Qz1, showing an approximately normal distribution, with just one high outlying value.

For example, for GFZ-Qz1 the 240 measurements showed an approximately normal distribution (Fig 7) with a mean of 0.078‰, a standard deviation of 0.013‰ and only one potential upward outlier of 0.127 ‰ with a z-score of 3.6. The fact that the mean value of 0.08 ‰ was constant across all of the RMs tested, suggests that it arises primarily from the instrument, partially as the counting statistics. However, at the typical count rate of 5 million counts per second for $^{18}O^-$ (and 2.5 billion cps for $^{16}O^{-)}$ and a total integration time of 80 seconds, one arrives at Poisson statistics limit of ~0.05 ‰. This value is calculated for the dominant source of uncertainty which is $^{18}O^-$, from the typical total number of counts (n) of 400 million (i.e. 5E06 * 80) and the predicted Poissonian standard deviation of 20,000 (√n), which gives a relative standard deviation of 0.05 ‰ (i.e. 1000*2E04/4E08) . This suggests that s_{inst} of 0.08 ‰ must also contain one or more components in addition to the Poisson statistics limit, such as sample charging, primary beam density changes, electron beam centering, electronics temperature, in addition to amplifier drift. Interestingly, gas source $\delta^{18}O$ determinations, which have orders of magnitude higher ion currents (i.e.,

a factor of 10 or more better than Poisson statistics) also never achieve Poisson results (Nasdala *et al*. 2016, Li *et al*. 2010 and Wiedenbeck *et al*. 2004). When the results from a single analysis is based on an integrated mean for the full 80 seconds of data acquisition then variations in secondary ion emission conditions (sample charging, primary ion flickering, etc.) over the course of the 80 s will, therefore, not directly influence the uncertainty estimates beyond any impact of an increasing or decreasing count rate during the 80 seconds.

Our assumption that there is negligible small-scale heterogeneity over 50-60 µm can be tested, albeit indirectly. A comparison between estimates of s_{meas} on the NIST 610 glass, show some apparent variability from 0.12, 0.11 and 0.10 ‰ by Method 2, and very similarly 0.12, 0.11 and 0.08 ‰ by Method 1 (for runs of the same 3 RMs, NBS-28, ZRM-1 and MfN-Qz1 in Table 2). We believe this glass most likely has the least heterogeneity at this spatial scale, in which case such observed variation may well reflect the improving stability of the instrument, particularly for MfN-Qz1 which was run a year later (Table 1). Over these same 3 runs, the estimates of s_{meas} made using the quartz RM (with Method 3) make a similar trend 0.14, 0.15 and 0.10 ‰, so this would confirm that this change is being driven by instrumental stability, rather than changes in small-scale heterogeneity. The somewhat higher estimates made by using the quartz candidate RM s (larger by around 0.02‰), could theoretically be the result of slight small-scale heterogeneity, but the fact that this additional value is quite similar for all three RMs would be more consistent with the explanation that it is due to a different behaviour between quartz vs. glass under the ion beam. If this is the case, then we have no evidence for any heterogeneity of $\delta^{18}O$ at the 50-60 µm scale in either the NIST-610 glass or in these quartz RM s.

Moreover, the fact that the values of s_{meas} decreased across these 3 runs, but s_{inst} stayed stable at 0.08‰, suggests that the improved stability of the instrumentation was not reflected at all by s_{inst}. This suggests that s_{meas}, estimated in this way when there is negligible heterogeneity at the 50 µm scale, gives a more meaningful value for the repeatability of the measurements, and should be preferred over either s_{inst} or Poisson statistics when operating at such high counting rates.

Outliers: comparison of manual identification vs. automatic accommodation via robust statistics

Taking the example of NBS-28 discussed above (Fig 4), the manual identification of outlying values resulted in the rejection of 3 excessively divergent individual measurements (based upon z-scores > 3), representing 1.5% of the 200 total values. Because they occurred in different duplicate pairs, this resulted in the rejection of 3 duplicates, which is 3% of the 100 pairs. Classical ANOVA on the 100 pairs was clearly affected by the 3 outlying pairs, as their removal reduced the value of s_{total} to 0.31‰ from a previously unreported value of from 0.36 ‰. The alternative option of using Robust ANOVA on all 100 pairs gave a similar value of 0.33‰. For the measurement repeatability, the manual removal lowered the classical estimate from to 0.14‰ from a previously unreported value of from 0.26 ‰., but this was the same as the value estimated by robust accommodation. This suggests that two approaches are equally effective.

The case of ZRM -1 is in some ways similar to that for NBS-28. Classical ANOVA on all 100 duplicate pairs of measurements (M ethod 3) gave an unusually large estimate of measurement repeatability of 0.22 ‰ (previously unreported), whereas robust ANOVA gave a significantly smaller value of 0.15 ‰. This difference between classical and robust estimates is very informative, indicating the presence of outlying values that require more detailed investigation. As reported in Table 1, this investigation revealed 9 pairs of craters with a separation distance of less than 50 μm, requiring rejection, A tenth pair was rejected on the basis of one analysis having a z-score of 3.3. When these 10 pairs were removed manually, classical ANOVA gave a measurement repeatability of 0.15 ‰, which effectively identical to that from robust ANOVA on all 100 raw pairs of values. (Table 2).

In certain respects, the situation for ZRM-1 is different when the variances related to heterogeneity are studied. The 10 ‘outliers’ have much less impact on the classical estimate of the heterogeneity ($s_{hetero[50um-bottle]}$) being 0.18 ‰ for n=90, as compared to 0.19 ‰ for n=100. Unexpectedly, robust ANOVA gives a higher result (0.21 ‰). Furthermore, the total variability of all of the measurements (s_{total}) for the raw data by classical ANOVA is 0.29 ‰ and reduced to 0.23 ‰ after manual filtering, but only

reduced to 0.25 ‰ using robust ANOVA on the raw values. Both of these effects highlight that robust ANOVA has difficulty in accommodating outlying values when the proportion of outlying values is at the specified limit (i.e. 10%, in this case 10 outlying pairs out of 100 pairs)

Overall, we therefore argue that the use of robust ANOVA is generally a reliable approach to identifying and estimating both repeatability and heterogeneity. The manual identification and exclusion of outlier values (Table 1) is equally effective, but it depends on a very thorough inspection of every measurement value. Usefully, robust statistics alerts the user to the possible presence of outlying values, which can then be investigated in more detail. However, all robust statistical programs do have the limitation that they cannot accommodat e more than a fixed proportion of outliers (in this case, 10%). It could be argued that such datasets are not suitable for heterogeneity estimation anyway, but the example of ZRM-1 suggests that manually eliminating measurement values arising from gross errors, such as too close or overlapping craters, can allow robust ANOVA to accommodate the typically rare occurrences of outlying values from other causes (as demonstrated by NMS-28).

Estimating the number of measurements needed to achieve a target uncertainty

Achieving a target analytical repeatability (or uncertainty) requires that the heterogeneity of a candidate RM is low enough to make this possible. For the example of $^{18}O^-/^{16}O^-$ by SIM S, if the target uncertainty (u′) is set at 0.11 ‰, and the measurement repeatability (s_{meas}) is 0.10 ‰, then the heterogeneity of the RM (s_{hetero}) must be not contribute more than 0.046 ‰, calculat ed using a rearrangement of Equation 1

$$s_{hetero} = \sqrt{s_{total}^2 - s_{meas}^2} = \sqrt{0.11^2 - 0.10^2}$$

Given the values of heterogeneity of candidate RMs measured in this study (the lowest of which is 0.18 ‰ for ZRM-1), it is clear that the uncertainty for a single measurement on any of these RMs will not be low enough to achieve this target uncertainty (i.e. u′ = 0.11 ‰).

One approach to achieving this target uncertainty would be to prepare a new RM that has a heterogeneity less than 0.046 ‰, but this would likely prove technologically infeasible. A second alternative approach, often adopted when heterogeneity is the dominant component of the total variance, is to take multiple measurements on different (randomly chosen) fragments of the RM and then use the mean of these '*n*' measurements. This is because taking the average of 'n' duplicate measurements made on 'n' pairs will reduce the effect of any heterogeneity on the mean value, whether it occurs within- or between- fragments. The uncertainty of this mean value (u_{mean}) can be estimated by the standard error on the mean value:

$$u_{mean} = s/\sqrt{n} \quad \text{.....(4)}$$

or

$$n = (s/u_{mean})^2$$

However, if the total variance were dominated by s_{meas}, even though this might include a small component from small-scale heterogeneity, then theoretically the uncertainty reduction could be achieved by making all 'n' measurements on a single fragment. This is an unlikely situation that was not found in any of the materials investigated during the current study. Furthermore, on materials produced with grains of only a few hundred µm there would not be sufficient space on a single grain to apply this strategy. For RMs where heterogeneity is the dominant source of the variance, as in this study, the best strategy will depend on the grain size of the RM. For fine-grained RMs the best strategy would be to have 'n' duplicate pairs of measurements, with each pair on 'n' randomly selected fragments. For course-grained RMs, the best strategy would be to have 'n' multiple pairs of measurements randomly distributed across each of the large fragments (the maximum number of which that can be fitted onto the sample mount).

Where each sampling event produces one measurement, as is the case for SIM S, the appropriate experiment al standard deviation to use is s_{total}, and the target value of the

uncertainty on the mean value (u'_{mean}) is estimated by the target standard error on the mean, we get

$n = (s_{total}/u'_{mean})^2$……..(5)

As an example, the NBS-28 quartz yielded a measured $s_{total} = 0.31‰$ which is dominated by heterogeneity (81% of total variance, Fig 5a).

If the target $u'_{mean} = 0.15$ ‰
the required number of randomly selected fragments of NBS-28 to be measured (n), from Equation (5) is therefore

$n = (0.31/0.15)^2 = 4.3 \approx 4$

Similarly, for a target uncertainty of 0.10‰ we have

$n = (0.31/0.10)^2 = 9.6 \approx 10$

The random selection of fragments is needed so as to minimize the risk of introducing bias, even when there is minimal heterogeneity between fragments. The certification procedure from the RM also aims to eliminate this bias, as discussed below.

Such numbers are totally feasible for SIMS work, where analytical experiments often provide many determinations on a calibration material during the course of a single day. When this calculation is applied to the other three materials investigated during this study, applying a target uncertainty of 0.10 ‰, the required numbers of fragments are 530, 5 and 7, for GFZ-Qz1, ZRM-1 and MfN-Qz1, respectively. All of these values assume that the heterogeneity property is uniform across the given material. This was clearly not the case for GFZ-Qz1, which is therefore unsuitable for further development as a $\delta^{18}O$ RM . For M fN-Qz1, which is currently provided as four large fragments (~5mm) from a single crystal, it would not be possible to mount 7 such fragments on a single 25 mm diameter block, let alone including with the test material to be investigated. In such cases as MfN-Qz1 theheterogeneity information could be

used to calculate the optimal grain size in which to supply a material to the user community. There was detectable extra variance between the current ~5 mm fragments, but it might be possible to produce smaller fragments (e.g. 250 μm) that have acceptably small levels of heterogeneity between them. This should give the same total variance (0.27 ‰), and hence the same number of 7 fragments required, but these could be fitted onto a single sample mount.

This approach has the interesting implication that the total uncertainty can potentially be made lower than the value of s_{meas}, by measuring enough fragments. In the case of NBS-28, a target uncertainty of 0.10‰ can be achieved on the mean of 9 fragments, which is well below the observed value of s_{meas} 0.13 ‰. This would imply that it is possible to quote an uncertainty on an assigned value of such a material, which could be lower than that for a bulk material, if you specify taking measurements on a large enough number of fragments. This assumes that any systematic effects, such as bias against the bulk value, have been corrected for, and the uncertainty of that correction included in the estimate of the measurement uncertainty. The uncertainty on the certified value would be based upon the standard error of the mean of 'n' measurements on 'n' fragments, rather than on the standard deviation of a single measurement taken on a single fragment. Here too we should stress that the total uncertainty from a bulk characterization has many additional sources of uncertainty – such as gas yield during decomposition, gas pressure in the ion source, etc. – which are outside the topic of the current research.

This approach suggests that in order to achieve a target uncertainty (u'_{mean}) of 0.10 ‰ a user would need to choose 10 fragments of NBS-28 selected at random and mount them on each block. All 10 fragments would then be analysed using a single measurement (e.g. for $\delta^{18}O$), from which the mean and the *s* can be calculated. The predicted value of s_{total} on the calculated mean value (standard error on the mean = $s/\sqrt{n}$), using Equation (4) is 0.10 ‰. This could be compared with the assigned value (using "bulk" analytical methods) and its uncertainty, to define the instrumental mass bias and its uncertainty during a given SIMS experiment al sequence.

It is already possible to include heterogeneity of the RM into the estimated standard measurement uncertainty (u) of a certified value of a CRM (Kane *at al.* 2003), using an equation that sums the variances (VAR):

$$u^2 = VAR(Y_{mean}/\sqrt{N}) + VAR_{inhom} + VAR_{bias} \ldots \text{Equation 6}$$

where the first term is effectively the standard error on the mean for the general variable Y of N contributing laboratories' mean data, the second term adds the inhomogeneities (i.e. heterogeneity, usually between-bottle) and the last term adds the effects of the between-method bias. The uncertainty thus includes these systematic effects of between-lab and between-method bias, in addition to the random effects, such as repeatability and heterogeneity. This equation could be adapted to include the variance caused by the heterogeneity when reduced by the use of 'n' fragments such as follows. Idealistically, participants in a certification exercise could report the mean value of measurements on the specified 'n' fragments of the RM. This would be expected to reduce the size of the first term in Equation 6, compared with the use of a single fragment by each participant. Morerealistically, the certified value would be established independently using bulk samples of the RM (e.g. $\delta^{18}O$ by gas source M S). The uncertainty on the certified value for a 'micro-beam' technique ($U_{CV\ beam}$) could be estimated (potentially by just a single elite SIM S lab) by including the extra variance caused by the heterogeneity at the specified micro-scale $U_{HET\ beam,}$ adjusted for the specified number of replicates 'n' using Equation 4, in a way analogous to that recently described for small beam PXRF (Rostron and Ramsey 2017).

$$U_{CV\ beam} = \sqrt{\left(U_{CV}{}^2 + U_{HET\ beam,}{}^2\right)}$$

In the subsequent use of the CRM for method validation, users would also need to make measurements on 'n' fragments and use the mean value to calculat e the bias against the certified value, and test whether this bias was statistically significant in the usual way (Linsinger 2010).

Use of heterogeneity values in assessing candidate materials

To decide whether any of these four test materials make a suitable RM for $\delta^{18}O$, the key issue is the estimated overall heterogeneity, as reported above. As these values are all greater than the proposed target uncertainty of 0.10 ‰, they are nominally unsuitable, but could be made so using measurements to be conducted on multiple fragments to reduce the effect of the heterogeneity, as described above. Excluded from further consideration is GFZ-Qz1, due to its excessive and uneven heterogeneity, as discussed above; all three other materials could be suitable assuming enough fragments were available. ZRM-1 and NBS-28 are both supplied as numerous small fragments (~250 μm) so the required number of fragments (6 or 10, respectively) could be easily fitted on an individual, 25mm diameter mount, alongside the actual test materials. As already suggested, reducing the size of the current fragments of MnN-Qz1 from ~5000 μm to ~250 μm, might make it suitable, after further testing.

More generally, heterogeneity is clearly an issue for the selection of potential RMs. The presence of appreciable zoning in the isotopic composition is a common source of such heterogeneity (Tracy, 1982), and should be precluded early in the selection process. However, a caveat needs to be added that the testing which is done must reflect the form in which the RM will be delivered to the end user. It makes little sense to start looking for crystallographic controlled isotopic zoning (e.g. MfN-QZ1) if the end user is going to get a bottle full of random 500 μm fragments.

Potential improvements in experimental design

Future studies of the heterogeneity of candidate μRMs could have improvements to their experimental design, based upon the lessons learnt in the current study.

1. Uniform design across all RMs, as far as possible given any differences in the fragment size between the RMs, which should include:
    a. Having duplicate measurements separated in space by a uniform distance (e.g. 50 μm), but separated in time by a randomly selected time interval (to capture an intermediate level of repeatability, after drift-correction)
    b. Placing drift monitoring (DM) measurements at regular intervals across the run on a sufficiently homogeneous area of both the crystal

RM itself, and also in an area of homogeneous glass RM inserted into the mount.

c. Having sufficient amount of material available on the final test mount so that the number of analytical pairs that are determined during heterogeneity assessment is much greater than the number of pairs needed to achieve a desired target uncertainty.

2. The initial test mount for assessing likely heterogeneity should, if possible, contain large piece of the candidate material (e.g., mm-size fragments). This will allow testing the heterogeneity of the material over intermediate sampling scales. If the candidate material looks promising then it should be crushed and sieved to a finer grain size (e.g. few 100's of µm) that optimize the material for future needs.

Estimating repeatability of measurements on test materials

This paper has focused on measurements made on reference materials, but has implications for measurements made on unknown test materials (TMs) as well. There is a strong case to argued that the repeatability of measurements on TMs should be estimated using the TM itself, rather than on µRM. For example, if the duplicate method were applied within- and between-multiple sets of fragments of the TM ($n \geq 8$), the resulting estimates of repeatability would reflect the measurement results of that particular TM (with its own within-fragment heterogeneity and perhaps chemically complex composition), rather than those of a probably more homogeneous and less complex CRM . Additional measurements on a matched µRM would be needed, if the potential bias of the measurements were to be required for an estimate of the uncertainty of the measurement values. However, in complex natural materials it is often genuine heterogeneity that the analyst seeks to establish. In all likelihood, the use of a natural test material to define its specific repeatability will often prove impractical.

Conclusions

We have shown that the duplicate method with ANOVA is capable of quantifying *in situ* heterogeneity of $\delta^{18}O$ in quartz at the ~400 picogram test portion mass. Unlike previous published methods for determining trace element

concentrations in glass at the nanogram scale, we use closely spaced duplicates placed across the RM (analysed at randomly assigned time separations) to estimate the measurement repeatability at a fine scale. This enables the use of ANOVA to subtract this measurement repeatability and thereby quantify the heterogeneity at larger scales. We used around 100 pairs of duplicate measurements, on around 100 grains of the find grained RMs (< 1mm) or across four fragments of the courser grained RMs (~ 2 mm), to quantify the heterogeneity. We consider that this number of 100 is sufficient to allow for some unevenness in the spatial distribution of any heterogeneity, being more than 10 times larger than the theoretical minimum of 8 (Lyn *et al.* 2007), but considerably less than the 577 measurements calculated by Harries (2014) to prove a maximum of 30% heterogeneity contribution to the total variance in a specific case of trace element analyses. We believe that our method will be equally applicable for other analytes (such as trace elements or other isotopic systems) in other minerals and materials, although this may require adaptation of the experimental design for a given specific case. Our results suggest that robust ANOVA is generally more reliable than classical ANOVA, if there are a limited number of outliers (for this software < 10%). However, if *all* outlying measurement are identified and removed manually then classical ANOVA can usually be relied upon. If there are more than 10% of outlying values, robust ANOVA results become unreliable. Outlier analyses identified by the robust approach should be investigat ed in order to exclude them as indicative of serious heterogeneity issues (e.g. mineral inclusions or alteration along fractures) which could exist at a small scale but which could bias bulk analytical data. Regardless of using robust statistics vs. manual identification, any SIMS result on a candidate RM that is identified as an outlier should be investigat ed to see if there is a clear justification for rejecting that data point as unreliable (i.e., overlapping craters, scratch on sample surface, etc.) or whether the data reveals a more fundamental problem with the candidate material itself.

In order to make a reliable estimate of the heterogeneity, repeatability arising from the measurement design (s_{meas}) needs to be small. The repeatability based on various methods tested was 0.08 - 0.16 ‰ in this study, and generally higher than the 'instrumental uncertainty' reported in the manufacturer's software which was constant at 0.08 ‰. The repeatability estimated using the duplicate method, therefore, better reflected the stability of the instrument for measurements on a particular test material

than the instrumental uncertainty, or the counting statistics which were also unaffected by the machine stability, as proposed by Fitzsimons *et al*. (2000). The instrumental uncertainty is a theoretical prediction assuming the ions are generated randomly with a Poisson distribution. It is therefore not an observed variability that can be affected by instrumental factors. Drift monitoring, and correction if the slope coefficient of the drift is found to be statistically significant, was found to be required in order to achieve a meaningful value of measurement repeatability. The material used for the drift monitoring needs to be sufficiently homogeneous in the area selected, and it is also preferable that it be matrix matched both chemically and mineralogically to the material being investigated. In this study, the silicate glass NIST-610 fulfilled the first criterion, with no detectable heterogeneity within the areas investigated. However, NIST-610 did not fulfil the second criterion. A sub-area of a fragment of two quartz RMs (MfN-Qz1, and GFZ-Qz1 – F4) seemed to fulfil both of these criteria. It is recommended that *both* types of drift monitor are used initially, (i.e. one that has been demonstrated to have very low heterogeneity, and one that is matrix matched) until the time where the impact of matrix matching in measurement repeatability is more fully understood.

The estimated overall heterogeneity of $\delta^{18}O$ in the four test materials ranged from 0.18 ‰ [for ZRM1] through 0.25 ‰ (for MfN-Qz1) and 0.28 ‰ (for NBS-28) to the grossly heterogeneous value of 2.3 ‰ (for GFZ-Qz1). Such values need to be attributed to a specified test portion mass (~400 pg), and size of fragments (~250 µm for NBS-28 & ZRM-1, and ~ 5000µm for MfN-Qz1 & GFZ-Qz1), and should be clearly stated on the final Certificate of Analysis. Also stated would be an uncertainty for the certified value, which includes a heterogeneity component for a stated microanalytical test portion mass. It would also be beneficial if a certificate provided guidance on the number of fragments of the CRM that need to be investigated in order to achieve a stated uncertainty, probably on the mean values of measurements on all of these fragments. Initial calculations have been made of the number of fragments that are required to achieve a target uncertainty of 0.10 ‰ for $\delta^{18}O$ on a mean value across those fragments. Assuming that the measurement repeatability remains at the current levels, the numbers of fragments required for the three feasible candidate RMs are 10, 5 and 7 for NBS-28, ZRM-1 and MfN-Qz1, respectively.

As modern technology is capable of rapidly producing highly precise results, reference materials producers need to take into consideration which microanalytical techniques need to be supported by a material under development. Our work has shown that even modest levels of heterogeneity, levels that would have gone undetected even a decade ago, are becoming a limitation to overall data quality, including our ability to estimate method repeatability. All else being equal, the production of materials with sufficient space to conduct at least 4 analyses for the intended analytical method are required. This should be balanced against the need to provide individual grains small enough to allow the desired specified number of grains (e.g., 7 fragments of MfN-Qz1) on a single sample mount while leaving sufficient space to include the "unknown" test materials which are the focus of an investigation. For a material intended for SIMS $\delta^{18}O$ determinations this would suggest that the starting material be produced with a grain size of roughly $250 < \phi < 500$ µm. For µRMs being developed for laser ablation studies a size closer to 1 mm would seem more appropriate.

Achieving full performance of any analytical technique (e.g. 0.10 ‰ for SIM S) will require the development of an RM with an acceptably low level of heterogeneity at the intended test portion mass. Even then, to achieve the quoted level of uncertainty on the certified value, a specified number of measurements will need to be made on a specified number of fragments of the RM. Where RMs are used with higher levels of heterogeneity, or with fewer than specified numbers of measurements and fragments, then both uncertainty and bias will be introduced into measurements made on subsequent samples. Both the analyst and the data user should be continually aware of this risk. Also not to be forgotten, the reliability of the assigned values based on bulk analytical methods must still be included when determining any absolute isotope ratio or element abundance values.

Further research will be required to develop a model for predicting sample heterogeneity versus test portion mass (and number and mass of fragments). This could be based on data obtained at picogram scale, but would enable extrapolation to larger masses of material. Such information should give important information that would bridge the gap between microanalytical and bulk analytical methods.

Acknowlegements

U. Dittmann produced the high quality sample mounts which were essential for this work. F. Couffignal was responsible for SIMS data collection and for assuring the excellent stability of the SIMS tool. We thank also Korth Kristallel GmbH (Altenholtz, Germany) and the Museum für Naturkunde Berlin for providing the starting crystals for GFZ-Qz1 and MfN-Qz1, respectively. Finally, we acknowledge the valuable suggestions made by two anonymous reviewers of this paper.

Electronic Supplement A1 for paper GGR0497

This supplement describes the detailed procedure of how a software program was adapted to interpret the measurements from the particular experiment al design used in this study.

The particular programme used was RANOVA (Version 1.0), (Analytical Methods Committee 2014; Rostron and Ramsey 2012), which runs as a macro with Microsoft Excel. This programme is designed specifically for a balanced experiment al design with two analytical measurements on both of two duplicate samples, taken from a series of 'n' sampling targets in a three-level, balanced design (2 x 2 x n).

In the experiment al design used in this study, for the fine-grained RMs, there are in effect single measurements, on two samples, on each of 'n' fragments (1 x 2 x n). To apply this program, it was therefore necessary to modify the data input, to enable the program to run. Each analytical measurement was replicated exactly, for both of the sample duplicates, to produce a data structure in the required of 2 x 2 x n format.

Interpretation of the output of RANOVA had to be modified to overcome the effect of this modified data input. The lowest level of variance, labelled as 'analytical' in the software, was inevitably zero, because the supposed 'analytical duplicates' were identical values. The second level of variance in the output, labelled as 'sampling' in the software output, provided the estimate of the measurement repeatability. The third level of variance, labelled 'between-target', gave the estimate of the heterogeneity at that scale. It has been shown by simulation, that this procedure of constructing zero analytical variance at the lower level, produces reliable estimates of variance at the higher levels.

References

Analytical Methods Committee (1989)
Robust Statistics - How Not to Reject Outliers. Part 1. Basic Concepts. **Analyst, 114**, 1693-1697.

AMC (2001)
Robust statistics: a method of coping with outliers. **Technical Brief 6,** Analytical Methods Committee, Royal Society of Chemistry. www.rsc.org/images/robust-statistics-technical-brief-6_tcm18-214850.pdf

Analytical Methods Committee (2014)
Software RANOVA
http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/

Ashwal L.D., Wiedenbeck M. and Torsvik T.H. (2017)
Archaean zircons in Miocene oceanic hotspot rocks establish ancient continental crust beneath Mauritius. **Nature Communications,** DOI: 10.1038/ncomms14086.

BAM (1991)
Zertifiziertes Referenzmat erial Reinstoff Nr.1, Certificat e of Analysis, Bundesanstalt für Materialforshung und –prüfung. 5 p.

Boyd F., Finger L. and Chayes F. (1967)
Computer reduction of electron probe data**. Carnegie Institution of Washington Yearbook, 67,** 21–215.

Eggins S.M. and Shelley J.M.G. (2002)
Compositional Heterogeneity in NIST SRM 610-617 Glasses, **Geostandards and Geoanalytical Research, 26/3**, 269–286.

Eiler J.M., Graham C and Valley J.W. (1997)
SIM S analysis of oxygen isotopes: M atrix effects in complex minerals and glasses. **Chemical Geology, 138/3-4**, 221-244.

Ellison S.L.R. (2015)
Homogeneity studies and ISO Guide 35:2006. **Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement, 20**, 519-531.

Fabrega C., Parcerisa D., Rossell, J.M., Gurenko A. and Franke C. (2017)
Predicting instrumental mass fractionation (IMF) of stable isotope SIMS analyses by response surface methodology (RSM). **Journal of Analytical Atomic spectrometry, 32/4**, 731-748, DOI: 10.1039/c6ja00397d.

Fitzsimons I.C.W., Harte B. and Clark R.M. (2000)
SIMS stable isotope measurement: counting statistics and analytical precision. **Mineralogical Magazine**, 64, 1, 59-83.

Gy P.M. (1979)
Sampling of Particulate M aterials – Theory and Practice. **Elsevier**, **(Amsterdam),** 431 p.

Harries D. (2014)
Homogeneity testing of microanalytical reference materials by electron probe microanalysis (EPMA). **Chemie der Erde - Geochemistry, 74**, 375-384.

Hartley M.E., Thordarson T., Taylor C., Fitton J.G. and EIMF (2012)

Evaluation of the effects of composition on the instrumental mass fractionation during SIMS oxygen isotope analyses of glasses. **Chemical Geology, 334**, 312–323.

Huber, P.J. (1981)
Robust Statistics, **Wiley**, New York, pp301

IAEA (2007)
Reference Sheet for Reference M aterials NBS 28 and NBS 30. **International Atomic Energy Agency** (Vienna, Austria) 5 p.

Ishimura T., Tsunogai U. and Nakagawa F. (2008)
Grain-scale heterogeneities in the stable carbon and oxygen isotopic compositions of the international standard calcite materials (NBS 19, NBS 18, IAEA-CO-1, and IAEA-CO-8). **Rapid Communications in Mass Spectrometry, 22/12**, 1925–1932 DOI:10:1002/rcm3571.

ISO (2006)
Guide 35: Reference materials – General and statistical principles for certification (3rd edition). **International Organization for Standardization (Geneva, Switzerland)**, 64 p.

Jochum K.P., Weis U., Stoll B., Kuzmin D., Yang Q., Raczek I., Jacob D.E., Stracke A., Birbaum K., Frick D.A., Günther D. and Enzweiler J. (2011)
Determination of Reference Values for NIST SRM 610–617 Glasses Following ISO Guidelines. **Geostandards and Geoanalytical Research, 35/4**, 397-429.

JCGM 200 (2008)
International Vocabulary of Metrology – basic and general concepts and associated terms (VIM, 3rd edition). **Joint Committee for Guides in Metrology (Sevres)**, 136 p.

Kane J.S., Potts P.J., Wiedenbeck M., Carignan J. and Wilson S. (2003)
International Association of Geoanalysts' protocol for the certification of geological and environmental reference materials. **Geostandards Newsletter: The Journal of Geostandards and Geoanalysis, 27,** 227-244.

Kita N.T., Ushikubo T., Fu B. and Valley J.W. (2009)
High precision SIMS oxygen isotope analysis and the effect of sample topography **Chemical Geology, 264**, 43–57

Li X.H., Long W.G., Li Q.L., Liu Y., Zheng Y.F., Yang Y.H., Chamberlain K.R., Wan D.F., Guo C.H., Wang X.C. and Tao H. (2010)
Penglai zircon megacrysts: A potential new working reference material for microbeam determination of Hf-O isotopes and U-Pb age. Geostandards and Geoanalytical Research, **34**, 117–134.

Linsinger T. (2010)

Comparison of a measurement result with the certified value, **European Commission - Joint Research Centre Institute for Reference Materials and Measurements (IRMM)** from http://www.erm-crm.org/ERM_products/application_notes/application_note_1/Pages/index.asp x

Lyn J.A., Ramsey M.H., Coad S., Damant A.P., Wood R. and Boon K.A. (2007)
The duplicate method of uncertainty estimation: are eight targets enough? **Analyst, 132**, 1147-1152. DOI: 10.1039/b702691a.

Nasdala L., Corfu F., Valley J.W., Spicuzza M.J., Wu F.-Y., Li Q.-L., Yang Y.-H., Fisher C., Münker C., Kennedy A.K., Reiners P.W., Kronz A., Wiedenbeck M., Wirth R., Chanmuang C., Zeug M., Váczi T., Norberg N., Häger T., Kröner A. and Hofmeister W. (2016)
Zircon M 127 – A homogeneous reference material for SIMS U–Pb geochronology combined with hafnium, oxygen and, potentially, lithium isotope analysis. **Geostandards and Geoanalytical Research, 40/4,** 457–475, doi: 10.1111/ggr.12123.

Pankhurst M.J., Walshaw R., and Morgan D.J. (2017)
M ajor Element Chemical Heterogeneity in Geo2 Olivine Microbeam Reference M aterial: A Spatial Approach to Quantifying Heterogeneity in Primary Reference M aterials. **Geostandards and Geoanalytical Research, 41/1**, 85-91. DOI: 10.1111/ggr.12134.

Ramsey M.H., Solomon-Wisdom G.O. and Argyraki A. (2013)
Evaluation of *in situ* heterogeneity of elements in solids: implications for analytical geochemistry. **Geostandards and Geoanalytical Research, 37/4**, 379-391, DOI: 10.1111/j.1751-908X.2013.00236.x.

Rollion-Bard C. and Marin-Carbonne J. (2011)
Determination of SIMS matrix effects on oxygen isotopic compositions in Carbonates. **Journal of Analytical and Atomic Spectrometry, 26**, 1285-1289.

Rostron P. and Ramsey M.H. (2012)
Cost effective, robust estimation of measurement uncertainty from sampling using unbalanced ANOVA. **Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement. 17/1**, 7-14, DOI: 10.1007/s00769-011-0846-2.

Rostron P. and Ramsey M.H. (2017)
Quantifying heterogeneity of small test portion masses of geological reference materials by PXRF: implications for uncertainty of reference values. **Geostandards and Geoanalytical Research, 41/3**, 359-473. DOI: 10.1111/ggr.12162

Seitz S., Baumgartner L.P., Bouvier A.-S., Putlitz B. and Vennemann T. (2016)
Quartz Reference M aterials for Oxygen Isotope Analysis by SIMS, **Geostandards and Geoanalytical Research, 41/1,** 69-75. DOI: 10.1111/ggr.12133

Thompson M. and Howarth R.J. (1980)
The frequency distribution of analytical error. **The Analyst, 105**, 1188-1195**,** DOI: 10.1039/AN9800501188

Tracy RJ (1982)
Compositional zoning and inclusions in metamorphic minerals. In: **Ferry JM (ed)** Characterization of metamorphism through mineral equilibria. **Rev Min 10**, Min Soc Amer, pp 335–397

Wiedenbeck M., Hanchar J.M., Peck W.H., Sylvester P., Valley J.W., Whitehouse M.J., Kronz A., Morishita Y., Nasdala L., Fiebig J., Franchi I., Girard J.P., Greenwood R.C., Hinton R., Kita N., Mason P.R.D., Norman M., Ogasawara M., Piccoli R., Rhede D., Satoh H., Schulz-Dobrick B., Skar O.,
Spicuzza M.J., Terada K., Tindle A., Togashi S., Vennemann T., Xie Q. and Zheng Y.F. (2004)
Further characterisation of the 91500 zircon crystal. **Geostandards and Geoanalytical Research, 28**, 9–39.